LA-UR-17-25196
Approved for public release; distribution is unlimited.

Title: Next Generation Infrastructure Plan FY18-FY22

Author(s): Bonnie, David John; Coulter, Susan K.; Hick, Jason Cody; Hollander, Brett Jason; Lueninghoener, Cory; Martinez, Jesse Edward; Mason, Michael A.; Montoya, David Richard; Montoya, Andrew J.; Randles, Timothy C.; Santos, Ben V.; Sena, Phillip A.; Vandenbusch, Tanya Marie; Velarde, Ron Ray

Intended for: Report

Issued: 2017-06-28

LOS ALAMOS NATIONAL LABORATORY

Next Generation Infrastructure Plan FY18-FY22

Prepared for: Advanced Simulation and Computing Program

Prepared by: David Bonnie, Susan Coulter, Jason Hick, Brett Hollander, Cory Lueninghoener, Jesse Martinez, Mike Mason, David Montoya, Andrew Montoya, Tim Randles, Ben Santos, Phil Sena, Tanya VandenBusch, Ron Velarde

June 1, 2017

EXECUTIVE SUMMARY

Objective

Define a 2-5 year plan that documents facilities, operations, networking, storage and user support efforts to concentrate and coordinate resources within the ASC FOUS Program.

Goals

1. Provide a 5-year compute roadmap
2. Estimate storage bandwidth and capacity needs to satisfy ASC compute requirements. Validate with historical storage usage data.
3. Estimate network bandwidth and capacity and desired architecture needs to satisfy ASC compute and storage requirements. Validate with historical network usage data.
4. Define plans for expanding user support for future systems.
5. Estimate power, cooling and other facility needs and document the plan to satisfy those requirements.
6. Define the role operations will take to support the expanded compute, storage and networking capabilities.
7. Define the planned approach to security, monitoring, and integration activities.

Outline

Provide drivers and central planning criteria for the major elements of the FOUS subprogram, and identify plans for the next 2-5 years to inform and coordinate the necessary resources. Discuss lifecycle and replacement/augmentation plans for key hardware. Determine storage capacity and bandwidth requirements. Estimate network port/architecture needs to support storage and compute. Provide user support and operations methodology to support exascale systems. Address needs for facility power, cooling, structural, and water for next five years.

PLANNING BUDGET

Desired budget by project

The figures in Table 1 represent a forecast by FOUS subprogram element in order to implement the scope identified in this plan. The amounts do not represent total funding for the subprogram elements, but focus on the acquisition costs necessary to implement new projects or refresh current capability.

Table 1: FOUS project requests for new funding (in $1K)

Project	FY18	FY19	FY20	FY21	FY22
Parallel Storage	$ 2,000[1]	$ 0	$ 2,000[2]	$ 500[3]	$ 2,000[4]
Deep Storage	$ 1,000[5]	$ 1,000[6]	$ 1,000[7]	$ 1,000[8]	$ 1,000[9]
Network	$ 1,500[10]	$ 100[11]	$ 0	$ 0	$ 0
Facility	$ 1,000[12]	$ 8,000[13]	$ 4,000[14]	$ 0	$ 7,000[15]
Platforms	$ 0[16]	$ 0	$ 0	$ 0	$ 0
Monitoring	$ 0	$ 0	$ 0	$ 125[17]	$ 0
Operations	$ 0	$ 0	$ 0	$ 0	$ 0
User Support	$ 0	$ 0	$ 0	$ 0	$ 0
Total	**$ 5,500**	**$ 9,100**	**$ 7,000**	**$ 1,125**	**$ 10,000**

1 Campaign storage capacity increase to 1yr of data retention.

2 Campaign replace gen1 hardware.

3 Replace homes/project storage hardware and solution.

4 Campaign storage capacity increase to 1.5yr of data retention.

5 Replace library infrastructure.

6 Replace library infrastructure.

7 Replace library infrastructure.

8 Replace library infrastructure.

9 Replace library infrastructure.

10 Replace Turquoise network firewall.

11 New linecards/expansion for Crossroads

12 Replace failing equipment in SCC (air compressor, BCU, humidifier).

13 Design, execute mechanical install for Crossroads, design electrical install for Crossroads.

14. Execute electrical install for Crossroads.

15 Replace SCC roof

16 Implement USRC and FOUS research to address future system software needs at scale without requiring additional budget.

17 Replace monitoring hardware on 5-year lifecycle.

5-YEAR PLATFORMS ROADMAP

The ASC Program expects to procure and deploy the following systems over the next five years. New computing platforms are main drivers for other elements in FOUS.

ASC platforms impacting FOUS infrastructure

For the ASC Program, CTS-1 deployments are expected to be complete by beginning of FY18. They represented a total of 5 computational systems (Fire, Ice, Snow, Hail and Lysander) and are connected to 3 different networks. Though the hardware is deployed and the systems are in production, significant effort is required to stabilize the new Omnipath (?) network, identify and communicate new settings to users to optimize performance and resolve power distribution issues.

Significant effort is required to improve the warm-water cooling capability and power distribution of the Strategic Computing Center (SCC) prior to the arrival of Crossroads (ATS-3) and the new CTS-2 systems. Crossroads is anticipated in the 4th Quarter (Q) of Fiscal Year (FY) 2020, and the first CTS-2 systems are planned to arrive 1Q FY21.

Planning for facility improvements will need to commence in FY18 in preparation for ATS-5 and CTS-3 systems that are expected in FY25.

PLATFORMS ENVIRONMENT

The FOUS platforms environment planning process is heavily influenced by the ASC Program's platform deliveries. Preparation for new platforms environments starts with involvement with Fast Forward, Design Forward, Path Forward, and other early system acquisition projects. Staff participate and interact in vendor meetings and presentations focused on new architectures, tools, and techniques that are likely to be useful for future systems to impact their design and support prior to their incorporation into new systems. The Tri-Lab Common Computing

Environment (CCE) meetings help plan and coordinate for changes to operating system, scheduler, and other Tri-Lab supported software. The Exascale Computing Project (ECP) is also impacting the platform environment by encouraging a new software stack that is expected to accelerate application performance significantly.

Focus over the next 2 to 5 years is on learning new ways to integrate and deploy future systems (CTS-2, ATS-3) especially with respect to a new focus on improving user workflows, improving speed and efficiency of integration efforts for new systems with an emphasis on testing, completing Continuous Security Monitoring (CSM) to rewrite organizations security plan and reduce time for security approvals, Dedicated System Time (DST) improved accuracy and thoroughness of system validation, establish a new partnership for SLURM scheduler features and improvements.

Over the course of the next 2-5 years, the intent is to gain a better understanding of application workflows, how to distribute capabilities in support of the users, and keep the application environment up-to-date. This includes optimization and focus on user integration and the impact to the user through DTSs. Effective system management practices are necessary to ensure efficiency and scalability, while reducing the footprint of the current operating systems by moving toward more advanced memory architectures.

The focus of this process includes platform environment planning and development, which includes the integration and deployment of the next generation systems; which include ATS-3 and CTS-2.

The objective is to maximize the capabilities of the system and application runtime environments by improving the use of application workflows and integrating new platforms in the current environment, while strengthening expertise knowledge and tools to support the following capabilities:

1. Schedule and run application workflows efficiently
2. Efficiently integrate new platforms into our environment
3. Efficiently operate the platforms
4. Provide an effective environment of application workflows
5. Provide expertise and tools to codes teams to optimize their workflows

Some areas that will need to be addressed in the next five years include:

1. Expanding our understanding of application workflows and their performance.
2. Define a new distributed support model across groups, Centers of Excellence and Code Teams.
3. Need to support and understand new levels of monitoring to improve runtime support.
4. Consider methods to reduce memory footprint of operating systems to optimize usage of new memory architectures.
5. Consider how to support new diversity and extensive application library needs beyond what the operating system provides.
6. Evaluate moving to a new industry-standard software stack.
7. Find ways to handle the shift from heavyweight operating systems to applications that provide more of their own libraries, workflows, and tools.

The Platforms Environment will explore using the Ultra Scale Research Center (USRC) for addressing many of the gaps and working directly with industry partners to address several of the focal points.

FACILITY REQUIREMENTS

Power, water, cooling, and structural capability improvements are essential to supporting future ASC systems. To accomplish its mission, the ASC Program funds facility operations and drives improvements in three data centers at LANL, the Strategic Computing Center (SCC, also known as the Nicholas Metropolis Center for Modeling and Simulation), the Laboratory Data Communications Complex (LDCC), and the Central Computing Facility (CCF).

Gaps in facility capabilities need to be identified about 8 years in advance of new compute systems in order to allow facility upgrade projects to be completed in time. Facility upgrade projects normally cost $10M or more and take about 5 years to complete. Facility projects result in a lasting capital improvement to the ASC Program enduring beyond any single compute system. Installation projects exist to connect new computational systems to facility infrastructure and involve one or more distinct Physical Infrastructure Integration (PII) projects. These PII projects normally commence 1-2 years in advance of the system arrival, cost under

$10M, and the equipment is designed for a particular computational system (e.g. requiring detailed specifications for each rack and overall system layout).

There are three major drivers for new HPC facilities projects:

1. New compute requirements. Density improvements in compute system design and system layout generally result in new floor loading, power and cooling requirements. These new requirements may require new volume/flow for water or air-cooling systems that require different diameters of pipes, or power densities that require new cables and breakers.
2. Efficiency improvements. HPC facilities personnel are constantly monitoring and assessing the efficiency of the power and cooling systems to improve function and reduce environmental impact.
3. Equipment lifecycle. Equipment is often replaced or upgraded through preventative or corrective maintenance operations. Effectiveness of the equipment is partially determined through failures.

Strategic Computing Center

Crossroads and most of the CTS-2 systems will be sited at the SCC. The Exascale Class Computer Cooling Equipment (EC3E) Project is increasing the warm-water cooling capability of the SCC to at least 8,000 tons (28.2 MW) by 2020. Facilities projects in the SCC normally range from $1M-50M and require support and management from LANL's Project Management Division. Specific plans for FY18 include changes to redistribute more power to the machine room. FY19 and FY20 will require design, execution and completion of the electrical and mechanical pieces of the Crossroads installation. FY22 will require replacement of the SCC roof.

Laboratory Data Communications Complex

The open computing CTS-2 system will likely be sited at the LDCC. ASC is supporting an HPC infrastructure project to help consolidate equipment out of the CCF into the LDCC as hardware is refreshed. Projects in the LDCC tend to be planned and executed internally because they normally range from $50K-$3M in cost due to the scale of the solutions in the facility.

Central Computing Facility

No new requirements exist and equipment should be consolidated into the LDCC or SCC to reduce costs and effort required by FOUS. Currently, the CCF houses testbeds and prototype systems for the ASC Program. Facilities projects in the CCF are almost exclusively led by FOD-UI or HPC facilities staff and normally do not exceed $1M due to the small scale of the equipment in the facility.

NETWORK REQUIREMENTS

Network Infrastructure planning depends largely on storage and compute needs. The network design assumes overall network infrastructure lifespan of ten years, currently ranging from FY15 to FY25. The network requires sustaining system connectivity for 3-5 years while enabling growth of at least 50% for future/unknown systems. The capability for retiring systems, enabling new systems, and upgrading/replacing select components through the whole network infrastructure is important to maintain. The biggest network driver is storage capacity and providing the full throughput for the file systems, campaign and archive transfer rates between each other as well as to the computer cluster/data transfer systems. Typically, requirements drive network infrastructure to be modular enough to add/remove capacity as needed. For example, increasing switch/line card count or interconnectivity for higher throughput needed.

Campus network

The current campus network requires sustainability of existing infrastructure over the next few years. A majority of the existing Ethernet backbone infrastructure for ASC systems is expected to last until mid FY19. It is expected that the archive will need to be replaced along with Crossroads coming in around FY20; we will need to expand out additional infrastructure to support these new systems starting in FY19. This will include additional switch and cable infrastructure. As designed, the Ethernet Chassis' for campus backbone will not need replacement until FY25.

Turquoise Firewall that ASC helps support and use will need an upgrade as soon as the FY18 timeframe. Our current plans include looking at replacing the firewall with a possibility of high

availability. Upgrading the existing firewall in FY18 will allow us to maintain until FY22 before requiring another refresh.

Storage network

Storage Area Network (SAN) infrastructure can currently accommodate 2-4 additional compute clusters and/or file systems depending on their expected sizes with the thought of needing to upgrade the network within the FY19-20 timeframe to accommodate newer technologies and systems that should last past FY22. The existing SAN infrastructure is primarily RDMA/InfiniBand based and can be utilized to include future upgrades/replacements of storage systems that may make use of RDMA/InfiniBand technology (Home/Projects, Scratch, Campaign, Archive).

External network

Existing DISCOM infrastructure is expected to stay at 10Gb/s due to hardware encryption limitation with the knowledge that 40/100Gb/s maybe on the horizon. Upgrade to DISCOM network requires collaboration with the Tri-Labs who make use of the network. Currently, there is no drive to expand capability above 10Gb/s. Focus for the next 2-5 years will be on monitoring the 100Gb/s technology being tested by DISCOM sites, and any increase in user demand.

STORAGE REQUIREMENTS

The driver for storage growth is the amount of total system memory available. Statistics show that users generate approximately 3x the total system memory per month or about 30x the total system memory per year. There is an overall focus for the next 2-5 years in multiple storage efforts (homes/project, scratch, campaign, and archive) to employ higher performing solutions that are nearest platforms and applications to better satisfy user demand, to establish a storage hardware refresh plan, and to change usage of the archive from a store and retain everything to an archive what is necessary usage model. Efforts to replace hardware and improve performance need to be coordinated with networking. The usage model needs user support

and will take new tools, methods, and constant reinforcement to modify. These changes should improve user satisfaction with storage while reducing overall storage costs.

Projected storage bandwidth

Homes & Project storage current bandwidth is adequate as provisioned. The storage is not designed to store large individual files or mass quantities of data into the system. User quotas on home and project spaces limit misuse and its broad effect on the shared storage system. It would benefit users to move homes/projects storage to the storage SAN (InfiniBand-based) for consolidation of network administration and optimizing latency for improved responsiveness.

There are dedicated scratch file systems purchased with each ATS system to prevent shared usage from impacting productivity on the largest compute systems. The scratch file system on Crossroads is expected to be about 40 TB/s. This bandwidth is based off a combination of system memory size and system Mean Time To Failure (MTTF) to avoid data loss due to platform hardware failures. Shared scratch file systems are employed on CTS systems and currently have a bandwidth capable of ingesting data at peak performance from the largest CTS system in production. Bandwidth on individual CTS systems is tunable by selecting the number of LNET routers or I/O nodes configured for each system. Current shared scratch file system bandwidth peak for well-formed I/O is over 100 GB/s. This current generation of hardware should be in service through 2021. Improvements to consider are using multiple metadata servers to increase metadata performance, and methods to increase IOPs and bandwidth.

Campaign storage is intended for storing data associated with each capability computing campaign (e.g. ATCC-1, ATCC-2, …). The goal is to reduce the frequency and amount of data archived during a computing campaign. The current bandwidth for campaign storage is 1 GB/s per PB. Metrics to better understand how campaign storage is used will help refine the bandwidth target. Campaign storage will likely require 1 TB/sec, accommodations for large data sets, and multi-TB sized single files in the 2020 timeframe. The storage in campaign will favor high bandwidth operations (large data sets with averaged sized files or a single large file).

Archival storage is being redesigned to favor recall bandwidth (as opposed to ingest bandwidth). Changes will begin to favor improved batch processing over interactive performance. The archival storage system needs a major technology refresh to improve performance and reduce cost. Improvements to allow user quotas to prevent misuse, features

such as parallel tape stripes to improve retrieval of multi-TB sized files within a 12-hour period are desired (batch expectations are overnight, requested by 6pm and available for I/O by 6am).

Anticipated capacity

Homes & Project will not need increased capacity in the next two or more years. Extra capacity is expected from enabling de-duplication and compression on existing storage. It is not expected that enabling these will have a negative impact on job performance, testing will validate before production use. De-duplication and compression will provide 180% increase in current capacity.

The current technology in use for Homes & Project is a major scalability and performance concern as Crossroads approaches. A study will collect and characterize mount stats from the cluster and filer perspective to project future demands. This information will be critical in determining the next generation of home/projects storage in 2020.

Dedicated scratch for Crossroads will likely require 100 PB of capacity. Increases in ATS system memory will require increases in dedicated scratch system capacity. Shared scratch is sized reasonably for CTS systems. Future CTS systems will need to provide shared scratch capacity increases in the event those future CTS systems are the largest systems using the shared scratch.

Campaign storage capacity is based on ingest of 3 total system memories per month and retention of the data for up to 1 year. To accommodate Trinity and CTS-1 platforms, campaign storage capacity should be 72 PB by the end of 2018. Crossroads memory is expected to be about 4 PB requiring campaign storage capacity of about 144 PB by 2020. Campaign storage initial quotas will be no larger than 70-80% of total system capacity. This will allow for growth and increase to quotas of unanticipated power users. Campaign storage will progress through hardware and technology refreshes in the growth to Crossroads and beyond towards exascale. These refreshes may impact user data residency. It is expected that users would have to copy any data they wish to be retained from the retiring version of campaign storage to the new version of campaign storage. It is expected that users will not archive all data in campaign storage, but be selective about which data is most important to retain. Users will be notified at least one month prior to retirement of campaign storage.

The duration of standard compute campaigns are around 6 months. At the end of this compute cycle data in Campaign Storage that corresponds to completed compute cycle will become read-only. This data will reside for 90 days and then will no longer be available through an expiration process. Prior to expiring, data should be reviewed to determine what is appropriate to save longer term to archival storage.

Archival storage capacity is essential to maintain, but it is a main focus over the next 2-5 years to reduce usage from storing everything to storing the most important data. This will require user support and engagement (training, policies) to effect change. Hardware refresh is essential to reducing cost and improving performance and reliability. Changing our approach to managing the archive technology as early adopters of new capacity will drive costs down and improve overall capability of the system. In replacing the hardware, emphasis will be given to reducing interactive usage and improving batch access to data. This will relax bandwidth demands on the archival storage system and should increase user satisfaction.

MONITORING

Monitoring is important for optimizing system, facilities, and security on CTS and ATS systems. Effort is focused on developing, deploying and maintaining the data collection and analysis systems. Coordination occurs with the CSSE area to perform machine learning or other advanced analysis of collected data. There is shared responsibility in selecting which analysis should be incorporated into production HPC operations or with system administrators (Platforms Environment) to improve production management of CTS and ATS systems. In general, analytics will be adopted as it is shown to improve mean time to repair or mean time to failure for CTS and ATS systems.

System Monitoring

This is everything we get directly from syslog that goes into Splunk. Data rates are 25-50 GB/Day. Users of the data are HPC Operations and system administrators in Platforms Environment. Drivers for system monitoring are new CTS or ATS systems or new infrastructure,

and hardware refresh on a 5-year cycle. Hardware is new in 2017 and will require a refresh in 2022.

Facilities Monitoring

Generally includes information related to the facilities, electricity, and water, environmental data like air humidity and temperature. Data rates are expected to be 10 GB/day. Users of the data are the Facilities staff, HPC Operations and CSSE machine learning analytics staff. The drivers for facilities monitoring are current infrastructure redesign, control system licenses used for proprietary monitoring and control, along with 5-year hardware refresh cycle. For the next 2-5 years, the current system for facilities monitoring will likely require a complete replacement and new architecture to capture the data required.

Enhanced Monitoring

This covers the collection and use of data not specifically included in System Monitoring and mostly used for a special purpose such as detailed diagnosis or tracing activity, some examples are:

1. LDMS
2. Darshan
3. SMART Data

Data rates can be very high ~4 TB/Day. The main users of the data are system administrators (Platforms Environment), user support and CSSE machine learning staff doing detailed debugging or problem diagnosis. Drivers for the enhanced monitoring are the storage (~4TB/Day) for this data, and requirements for the Data Analytics Cluster (DAC). Focus over the next 2-5 years in this area will be developing and deploying the new Data Analytics Cluster, and completion of the Continuous Security Monitoring (CSM) solution to improve timeliness of security approvals for new systems.

OPERATIONS

Operating supercomputers for the ASC Program is an essential around-the-clock endeavor, and is critical to optimizing our system(s) availability and milestones. As such, operations is mostly effort-based and improvements center around enriching knowledge, skills, and abilities of staff. The overall goal of operations is to focus on activities that increase mean time to failure, reduce mean time to repair and improve proactive responses to problem diagnosis to ease the burden on deeper levels of support.

Operation Support

The team is known as Tech Ops and consists of 3 different shifts that cover weekdays/weekends/Holidays and the winter closure. The team currently manages spare parts logistics with vendors; this starts with systems arrival and continues through to decommissioning. The team also serves as a key point-of-contact for vendor support contracts handling round-the-clock communication and coordination for problem resolution. The majority of hardware issues are identified and resolved by members of this team. Over the last few years the Tech Ops Staff has taken on more administrative tasks. For example, there are several members of the team that assist in the weekly on-call duties for the archival storage and backup systems. The team also coordinates some outages. Increasingly, hiring people with advanced knowledge of compute and storage systems has improved responses to and resolution of hardware issues in those areas. The staff has an excellent track record of helping to perform first response administration and maintenance of infrastructure and computational systems. The team provides monitoring support for the D-WAVE System, they also assist the facility team by filling the machine with nitrogen when needed, and help escort the D-WAVE employees as needed. Establishing and maintaining procedures is especially important when dealing with system failures and the team utilizes wiki pages, and creates new documentation. Tech Ops participates in decommissioning systems. The focus for the next 2-5 years will be on continuing to improve knowledge, skills, and abilities by improving system availability.

Triage Team

In order to further improve availability, Operations is currently making a transition to a tiered support model. From the operations team we have decided to develop a new team that will work more closely with system administrators to improve skills for career growth. The specific focus of this team for the next 2-5 years is in reducing mean time to repair for hardware failures in ATS, CTS, or infrastructure systems (e.g. storage, facilities, networking). This will also enable hiring of individuals with specific HPC experience into operations creating a pipeline of talent. The team is expected to have broad knowledge of HPC systems and be able to handle the majority of tickets that are created by our users to resolve as many as possible quickly. This will allow other staff to focus on more complex issues, reduce the time to respond to user issues, and improve user satisfaction.

Hardware Support Tracking

Supports testbed development, accomplishes the relocation of computer racks, and cleans computer rooms to maintain a satisfactory computing and storage environment. Another function of this team is to research new technology.

USER SUPPORT

Providing quality technical direction and support to customers is critical to accomplishing high performance computing. Areas of responsibility include understanding production computing, providing user documentation, necessary training, account processing, user communications and maintain the required services, applications, and tools for internal operations. Drivers for user support are new computing environments that arise from hardware or software changes, and new scientific areas using computing systems. Individual team member expertise varies, but in general there is a breadth of knowledge in various areas and topics ranging from physics, fluid dynamics, computer science, and programming. Besides providing phone support (user help desk), the team provides compilation assistance, storage expertise, workload management advice, visualization support, and general customer service to help users and system administrators collaborate rather than terminate.

Focus for the next 2-5 years will be on timely initial response to problems, ensuring communication with users when issues endure, and helping to modify usage of ATS and CTS systems to match evolving design/architecture.

While every issue varies, the user support team is staffed to handle short to mid-term issues ranging from a few minutes to a few months. Metrics on responsiveness, characterizations of issues, and ultimate resolution of issues will provide insight and establish goals for improvement.